Analysis and Applications

LEARNING RATES FOR DENSITY LEVEL DETECTION

CLINT SCOVEL, DON HUSH, and INGO STEINWART
Modeling, Algorithms and Informatics Group, CCS-3
Los Alamos National Laboratory
Los Alamos, New Mexico, 87545
United States of America
{*jcs,dhush,ingo*}*@lanl.gov*

Received (April 6, 2005)
Revised (June 8, 2005)

In this paper we address learning rates for the density level detection (DLD) problem. We begin by proving a "No Free Lunch Theorem" showing that rates cannot be obtained in general. Then we apply a recently established classification framework to obtain rates for DLD support vector machines under mild assumptions on the density.

Keywords: anomaly detection; learning theory; rates; density level detection.

Mathematics Subject Classification 2000: 68Q32, 68T05

1. Introduction

This paper is a follow up to a recent paper [1] where we developed a classification framework for the density level detection (DLD) problem. Here we utilize recent results on classification from [2,3] to provide rate theorems for SVMs for the DLD problem. Let us begin by defining the density level detection problem. Let $(X, \mathcal{A})$ be a measurable space and μ a *known* distribution on $(X, \mathcal{A})$. Furthermore, let Q be an *unknown* distribution on $(X, \mathcal{A})$ which has an *unknown* density h with respect to μ, i.e. $dQ = hd\mu$. Given a $\rho > 0$ the set $\{h > \rho\}$ is called the *ρ-level set* of the density h. As in many other papers (see e.g. [4, 5]) we assume that $\{h = \rho\}$ is a μ-zero set and hence it is also a Q-zero set. Now, the goal of the DLD problem is to find an estimate of the ρ-level set of h. To this end we need some information which in our case is given to us by a training set $T = (x_1, \ldots, x_n) \in X^n$. We will assume in the following that T is i.i.d. drawn from Q. With the help of T a DLD algorithm constructs a function $f_T : X \to \mathbb{R}$ for which the set $\{f_T > 0\}$ is an estimate of the ρ-level set $\{h > \rho\}$. Since in general $\{f_T > 0\}$ does not exactly coincide with $\{h > \rho\}$ we need a *performance measure* which describes how well $\{f_T > 0\}$ approximates the set $\{h > \rho\}$. Probably the best known performance measure (see e.g. [5, 6] and the references therein) for measurable functions $f : X \to \mathbb{R}$ is

$$\mathcal{S}_{\mu,h,\rho}(f) \;:=\; \mu\Big(\{f > 0\} \,\triangle\, \{h > \rho\}\Big)\,,$$

where Δ denotes the symmetric difference. Then the goal of the DLD problem is to find f_T such that $\mathcal{S}_{\mu,h,\rho}(f_T)$ is close to zero.

The DLD problem is a well known problem in statistics and has important applications in anomaly detection (see e.g. [1,7] and the references therein) and many other areas. For example, it can be used for the problem of cluster analysis as described in [8,9] and for testing of multimodality (see e.g. [10,11]). Some other applications including estimation of non-linear functionals of densities, density estimation, regression analysis and spectral analysis are briefly described in [4].

In the statistical literature the most common approach for the DLD problem is the *excess mass* approach (see e.g. [12,10,4,5], and the references therein). Unfortunately this approach is based on empirical risk minimization and hence in general we cannot expect this approach to be computationally feasible (see however [12] for an algorithm with $\mathcal{O}(n^2)$ space and $\mathcal{O}(n^3)$ time requirements for a very special class of distributions on $\mathbb{R}^2$). To overcome this problem a method has been proposed in [1,7] that utilizes a classification performance risk for which quantitative comparisons with $\mathcal{S}_{\mu,h,\rho}$ can be achieved (see Theorem 3.1). This classification approach suggests efficient algorithms which will work for large classes of distributional assumptions. Indeed, in [1] an SVM is specified and universal consistency with respect to $\mathcal{S}_{\mu,h,\rho}$ proved.

In this paper we continue our investigation into the DLD problem by proving a "No Free Lunch Theorem". In addition, we use modifications of recent results of [2,3] applied to this classification framework to provide a learning rate theorem for the DLD problem in terms of a modification of the geometric noise exponent $\alpha \in (0,\infty]$ introduced in [2] and the noise exponent $q \in [0,\infty]$ introduced by Polonik [4]. That is, we show that the SVMs introduced in [1] obtain learning rates for $\mathcal{S}_{\mu,h,\rho}$ essentially of the form

$$n^{-\frac{q\alpha}{(1+q)(2\alpha+1)}}$$

if $\alpha \leq \frac{q+2}{2q}$ and

$$n^{-\frac{2q\alpha}{2\alpha(q+2)+3q+4}}$$

if $\alpha > \frac{q+2}{2q}$. A simple version of these rate results has already been announced in [7].

2. Definitions and Results

In this section we define terms and state our results. We begin by recalling the definition of noise exponent for DLD from [1]:

Definition 2.1. Let μ be a distribution on X and $h : X \to [0,\infty)$ be a measurable function with $\int h d\mu = 1$, i.e. h is a density with respect to μ. For $\rho > 0$ and

$0 \leq q \leq \infty$ we say that h has *ρ-exponent q* if there exists a constant $C > 0$ such that for all sufficiently small $t > 0$ we have

$$\mu\big(\{|h - \rho| \leq t\}\big) \;\leq\; Ct^q\,. \tag{2.1}$$

Definition 2.1 was first considered in [4, p. 864] where examples of distributions with ρ-exponent 1 for all ρ and examples with ρ-exponent $\frac{1}{2}$ for all ρ were described. In [1] this condition was shown to be closely related to a concept in binary classification called the Tsybakov noise exponent (see e.g. [5]).

We can now proceed to a No Free Lunch Theorem in the spirit of the well known result [13, Theorem 7.2] of Devroye et al. Note that here there is a conceptual difference since the density level ρ is often considered a tuning parameter and therefore it is desirable that a No Free Lunch Theorem for DLD guarantees the existence of densities for which detecting *all* of their density levels is hard. Such a result is provided by the following theorem proven in Section 3:

Theorem 2.1. *Let $\frac{1}{16} \geq a_1 \geq a_2 \geq \ldots$ be a strictly positive, decreasing sequence converging to 0 and μ be a measure on X which has no atoms. Then for every DLD algorithm $\mathcal{D} : T \mapsto f_T$ there exists a measure Q with density $h : X \to [0, 3]$ which has ρ-exponent ∞ for all $0 < \rho < \|h\|_\infty$ such that*

$$\mathbb{E}_{T\sim Q^n} \mathcal{S}_{\mu,h,\rho}(f_T) \geq \frac{3}{10} a_n$$

for all n and all $0 < \rho < \|h\|_\infty$.

Theorem 2.1 shows that learning rates are impossible without some restrictions on the distributions involved. To define such restrictions we consider a modification of the geometric noise exponent introduced in [2] for the classification problem. To that end we define

$$\tau_x \;:=\; \begin{cases} d\big(x, \{h > \rho\}\big) & \text{if } x \in \{h < \rho\} \\ d\big(x, \{h < \rho\}\big) & \text{if } x \in \{h \geq \rho\} \end{cases}$$

where d is the usual distance from a point to a set in the Euclidian space $\mathbb{R}^d$. We then define the *geometric* noise exponent as follows.

Definition 2.2. Let μ be a distribution on $X \subset \mathbb{R}^d$ and $h : X \to [0, \infty)$ be a measurable function with $\int h d\mu = 1$, i.e. h is a density with respect to μ. For $\rho > 0$ and $\alpha \in (0, \infty]$ we say that h has *geometric ρ-exponent α* if

$$\int_X \tau_x^{-\alpha d} |h - \rho| d\mu < \infty.$$

The exponent α describes the concentration of the measure $|h - \rho|d\mu$ near the set $\{h = \rho\}$ and does not imply any smoothness of the function h or the set $\{h = \rho\}$. However, one can show as in [2, Theorem 2.6] that if h has noise exponent q and h satisfies the envelope condition

$$|h(x) - \rho| \leq c_\gamma \tau_x^\gamma, \quad x \in X$$

for some constants γ and c_γ, then h has geometric ρ-exponent $\alpha = \frac{q+1}{d}\gamma$ if $q \geq 1$ and geometric ρ-exponent α for all $\alpha < \frac{q+1}{d}\gamma$ otherwise.

We now introduce the learning algorithms we will investigate. To this end let $k : X \times X \to \mathbb{R}$ be a positive definite kernel with reproducing kernel Hilbert space (RKHS) H. Let $l : Y \times \mathbb{R} \to [0, \infty)$ be the *hinge* loss function, i.e. $l(y,t) := \max\{0, 1 - yt\}$, $y \in Y$, $t \in \mathbb{R}$. Then for training sets $T^+ = (x_1, \ldots, x_{n_+}) \in X^{n_+}$ and $T^- = (x_1, \ldots, x_{n_-}) \in X^{n_-}$, a regularization parameter $\lambda > 0$, and $\rho > 0$ we define $f_{T^+,T^-,\lambda}$ to be a minimizer in

$$\arg\min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{(1+\rho)n_+} \sum_{i=1}^{n_+} l\big(1, f(x_i)\big) + \frac{\rho}{(1+\rho)n_-} \sum_{j=1}^{n_-} l\big(-1, f(x_j)\big)\,, \quad (2.2)$$

and $(\tilde{f}_{T^+,T^-,\lambda}, \tilde{b}_{T^+,T^-,\lambda})$ to be a mininizer in

$$\arg\min_{\substack{f \in H \\ b \in \mathbb{R}}} \lambda \|f\|_H^2 + \frac{1}{(1+\rho)n_+} \sum_{i=1}^{n_+} l\big(1, f(x_i)+b\big) + \frac{\rho}{(1+\rho)n_-} \sum_{j=1}^{n_-} l\big(-1, f(x_j)+b\big). \quad (2.3)$$

The decision function of the *SVM without offset* is $f_{T^+,T^-,\lambda} : X \to \mathbb{R}$ and analogously, the *SVM with offset* has the decision function $\tilde{f}_{T^+,T^-,\lambda} + \tilde{b}_{T^+,T^-,\lambda} : X \to \mathbb{R}$.

We can now state our main result which considers the sample plan where independently $n_+ = nm_+$ samples are taken *i.i.d.* from Q and $n_- = nm_-$ samples are taken *i.i.d.* from μ.

Theorem 2.2. *Let X be the closed unit ball of the Euclidian space $\mathbb{R}^d$, and μ and Q be distributions on X such that $dQ = hd\mu$ for some non-negative function h. For fixed $\rho > 0$ assume that the density h has both ρ-exponent $q \in [0, \infty]$ and geometric ρ-exponent $\alpha \in (0, \infty)$. We define*

$$\lambda_n := \begin{cases} n^{-\frac{\alpha+1}{2\alpha+1}} & \text{if } \alpha \leq \frac{q+2}{2q} \\ n^{-\frac{2(\alpha+1)(q+1)}{2\alpha(q+2)+3q+4}} & \text{otherwise}, \end{cases}$$

and $\sigma_n := \lambda_n^{-\frac{1}{(\alpha+1)d}}$ in both cases. Then for all $\varepsilon > 0$ there exists a constant $C > 0$ such that for all $x \geq 1$, $n \geq 1$, $m_+ \geq 1$ and $m_- \geq 1$, the SVM defined in line (2.2) using λ_n and Gaussian RBF kernel $k_{\sigma_n}(x, x') = \exp(-\sigma_n^2 \|x - x'\|_2^2), x, x' \in X$, satisfies

$$(Q^{nm_+} \otimes \mu^{nm_-})^*\Big((T^+, T^-) : \mathcal{S}_P(f_{T^+,T^-,\lambda_n}) \leq Cx^2 n^{-\frac{q\alpha}{(1+q)(2\alpha+1)}+\varepsilon}\Big) \;\geq\; 1 - e^{-x}$$

if $\alpha \leq \frac{q+2}{2q}$ and

$$(Q^{nm_+} \otimes \mu^{nm_-})^*\Big((T^+, T^-) : \mathcal{S}_P(f_{T^+,T^-,\lambda_n}) \leq Cx^2 n^{-\frac{2q\alpha}{2\alpha(q+2)+3q+4}+\varepsilon}\Big) \;\geq\; 1 - e^{-x}$$

otherwise. If $\alpha = \infty$ the latter concentration inequality holds if $\sigma_n = \sigma$ is a constant with $\sigma > 2\sqrt{d}$. Furthermore, all results hold for the SVM with offset defined in line (2.3) if $q > 0$. Finally, the notation $(Q^{nm_+} \otimes \mu^{nm_-})^*$ denotes the outer probability of $Q^{nm_+} \otimes \mu^{nm_-}$ and is used to avoid measurability considerations.

Remark 2.1. In the proof of Theorem 2.2 we prove rates for a classification risk $\mathcal{R}_P$ (see line (3.2)) which, although it may appear merely as a technical device, can be construed as a performance measure for density level detection with as much validity as $\mathcal{S}_P$. See [1] for a discussion.

Remark 2.2. (SVMs using Sobolev spaces) Theorem 2.2 is modeled on [2, Theorem 2.8] and chooses the Gaussian RBF parameter σ to depend on n and both noise exponents. However [3, Example 1] shows how results similar to [2, Theorem 2.8] can be obtained for classification with a fixed choice of Sobolev space for the RKHS. Using the same techniques we use to prove Theorem 2.2 we obtain the analogue of [3, Theorem 1] and therefore the analogue of [3, Example 1] for density level detection. The latter can be stated as follows: Let X be the unit ball in $\mathbb{R}^d$ and choose as a RKHS the Sobolev space $W^m(X)$ with $m > d/2$. Let μ and Q be distributions on X such that $dQ = hd\mu$. For fixed $\rho > 0$ assume that the density h has both ρ-exponent $q \in [0, \infty]$ and geometric ρ-exponent $\alpha \in (0, \infty)$. Then with the appropriate choice of regularization parameter we obtain optimal rates essentially of the form

$$n^{-\frac{4\alpha dmq}{(2mq+dq+4m)(2\alpha d+d+2m)}}.$$

3. Proofs

In this section we prove Theorems 2.1 and 2.2.

Proof of Theorem 2.1. The proof uses ideas from Devroye et al. [13, Theorem 7.2]. Let us first assume that $a_1 \leq \frac{1}{16}$ and define $\hat{a}_n := 2a_n$. If $(\hat{p}_n)$ denotes the sequence of [13, Lem. 7.1] with respect to $(\hat{a}_n)$ we write $p_n := \frac{\hat{p}_n}{2}$. Now recall that Lyapunov's theorem states that the image of every atom-free finite measure is a closed interval. Therefore, we can inductively find a partition $A_{-1}, A_0, A_1, A_2, \ldots$ of X with $\mu(A_{-1}) = \frac{1}{6}$, $\mu(A_0) = \frac{1}{3}$, and $\mu(A_n) = p_n$ for $n \geq 1$. Furthermore, let $\hat{\nu}$ be the measure on $\{0,1\}$ which is defined by $\hat{\nu}(\{0\}) = \frac{1}{2}$. We will use the product measure $\nu := \bigotimes_1^\infty \hat{\nu}$ on $\Omega := \{0,1\}^\infty$ for constructing "random densities". To this end we write $c_\omega := \frac{1}{3} + \sum_{i=1}^\infty \omega_i p_i$ for all $\omega = (\omega_i) \in \Omega$. Now, given an $\omega \in \Omega$ we define a density $h_\omega : X \to [0,3]$ by $h_\omega :\equiv 0$ on A_{-1}, $h_\omega :\equiv \frac{1}{c_\omega}$ on A_0, and $h_\omega(n) :\equiv \frac{\omega_n}{c_\omega}$ on A_n for $n \geq 1$. It follows that $\|h\|_\infty = \frac{1}{c_\omega}$. Consequently, the relation $s = \frac{1}{1+\rho}$ implies we only need to consider the s interval $(\frac{c_\omega}{c_\omega+1}, 1)$. Consider the shorthand notation $\mathcal{S}_{\omega,s} := \mathcal{S}_{\mu,h_\omega,\rho}$ where $\rho = \frac{1-s}{s}$ and let us fix an $s \in (\frac{c_\omega}{c_\omega+1}, 1)$ with the corresponding $\rho = \frac{1-s}{s}$. Since the definition of h_ω implies that $\{h_\omega > \rho'\} = \{h_\omega > \rho\}$ for all $\rho' \in (0, \frac{1}{c_\omega})$ and all $\omega \in \Omega$, denoting $s' = \frac{1}{1+\rho'}$, we obtain for any f that

$$\mathcal{S}_{\omega,s'}(f) = \mu\Big(\{f > 0\} \triangle \{h_\omega > \rho'\}\Big) = \mu\Big(\{f > 0\} \triangle \{h_\omega > \rho\}\Big) = \mathcal{S}_{\omega,s}(f).$$

Consequently any ω found to provide the inequality of the theorem for our fixed $s \in (\frac{c_\omega}{c_\omega+1}, 1)$ also works for any other value $s' \in (\frac{c_\omega}{c_\omega+1}, 1)$.

Now, for $T = (x_i) \in X^\infty$ we write $T_n := (x_1, \ldots, x_n)$ and obtain

$$\begin{aligned}
&\int_\Omega \inf_{n\geq 1} \frac{1}{a_n} \int_{X^n} \mathcal{S}_{\omega,s}(f_{T_n})\, Q_\omega^n(dT)\nu(d\omega) \\
&\geq \int_\Omega \int_{X^\infty} \inf_{n\geq 1} \frac{\mathcal{S}_{\omega,s}(f_{T_n})}{a_n}\, Q_\omega^\infty(dT)\nu(d\omega) \\
&\geq \int_\Omega \int_{X^\infty} \mathbf{1}_{\bigcap_{n=1}^\infty \{\mathcal{S}_{\omega,s}(f_{T_n}) \geq a_n\}} Q_\omega^\infty(dT)\nu(d\omega) \\
&\geq 1 - \sum_{n=1}^\infty \int_\Omega \int_{X^\infty} \mathbf{1}_{\{\mathcal{S}_{\omega,s}(f_{T_n}) < a_n\}} Q_\omega^\infty(dT)\nu(d\omega)\,.
\end{aligned}$$

Furthermore, for $\omega \in \Omega$, $i \geq 1$ and a decision function $f : X \to Y$ we write

$$E_{\omega,i}(f) := A_i \cap \Big(\{f > 0\} \vartriangle \{h_\omega > \rho\}\Big).$$

Now for $i \geq 1$ we define $\hat{f}_{T_n}(i) := \arg\max_y \mu(\{f_{T_n} = y\} \cap A_i)$, where in the presence of a tie we set $\hat{f}_{T_n}(i) := 1$. This definition implies

$$\mathbf{1}_{\{\mu(E_{\omega,i}(f_{T_n})) \geq p_i/2\}} \;\geq\; \mathbf{1}_{\left\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\right\}}\,,$$

and since $\mu(E_{\omega,i}(f)) \geq \frac{p_i}{2}\mathbf{1}_{\{\mu(E_{\omega,i}(f)) \geq p_i/2\}}$ we obtain

$$\mathcal{S}_{\omega,s}(f_{T_n}) = \mu\Big(\bigcup_{i=-1}^\infty E_{\omega,i}(f_{T_n})\Big) \;\geq\; \sum_{i=1}^\infty \mu\big(E_{\omega,i}(f_{T_n})\big) \;\geq\; \frac{1}{2}\sum_{i=1}^\infty \mathbf{1}_{\left\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\right\}} p_i\,.$$

Since this shows

$$\Big\{(\omega, T) : \mathcal{S}_{\omega,s}(f_{T_n}) < a_n\Big\} \subset \Big\{(\omega, T) : \sum_{i=1}^\infty \mathbf{1}_{\left\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\right\}} p_i < 2a_n\Big\}$$

we obtain

$$\begin{aligned}
&\int_\Omega \int_{X^\infty} \mathbf{1}_{\{\mathcal{S}_{\omega,s}(f_{T_n}) < a_n\}} Q_\omega^\infty(dT)\nu(d\omega) \\
&\leq \int_\Omega \int_{X^\infty} \mathbf{1}_{\left\{\sum_{i=1}^\infty \mathbf{1}_{\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\}} p_i < 2a_n\right\}} Q_\omega^\infty(dT)\nu(d\omega) \\
&= \int_\Omega \int_{X^n} \mathbf{1}_{\left\{\sum_{i=1}^\infty \mathbf{1}_{\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\}} p_i < 2a_n\right\}} Q_\omega^n(dT)\nu(d\omega) \\
&= \int_\Omega \int_{X^n} \mathbf{1}_{\left\{\sum_{i=1}^\infty \mathbf{1}_{\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\}} p_i < 2a_n\right\}} \prod_{j=1}^n h_\omega(x_j)\mu^n(dT)\nu(d\omega) \\
&\leq 3^n \int_\Omega \int_{X^n} \mathbf{1}_{\left\{\sum_{i=1}^\infty \mathbf{1}_{\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\}} p_i < 2a_n\right\}} \mu^n(dT)\nu(d\omega) \\
&\leq 3^n \int_{X^n} \int_\Omega \mathbf{1}_{\left\{\sum_{i \notin T} \mathbf{1}_{\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\}} p_i < 2a_n\right\}} \nu(d\omega)\mu^n(dT)\,,
\end{aligned}$$

where $i \notin T = (x_1, \ldots, x_n)$ means $\{i : A_i \cap \{x_1, \ldots, x_n\} = \emptyset\}$. Now, for fixed $T = (x_1, \ldots, x_n) \in X^n$ we denote by Ω_T the product of the x_ith components of Ω, $i = 1, \ldots, n$. Analogously, $\Omega_{\neg T}$ denotes the product of the remaining components

of Ω. Obviously we have $\Omega = \Omega_{\neg T} \times \Omega_T$. Analogously, the measure ν can be decomposed into $\nu = \nu_{\neg T} \otimes \nu_T$. With this notation we obtain

$$\begin{aligned} & \int_{\Omega} \mathbf{1}_{\left\{\sum_{i \notin T} \mathbf{1}_{\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\}} p_i < 2a_n\right\}} \nu(d\omega) \\ = & \int_{\Omega_T} \int_{\Omega_{\neg T}} \mathbf{1}_{\left\{\sum_{i \notin T} \mathbf{1}_{\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\}} p_i < 2a_n\right\}} \nu_{\neg T}(d\omega_{\neg T}) \nu_T(d\omega_T) \,. \end{aligned}$$

Now, we observe

$$\begin{aligned} \int_{\Omega_{\neg T}} \mathbf{1}_{\left\{\sum_{i \notin T} \mathbf{1}_{\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\}} p_i < 2a_n\right\}} \nu_{\neg T}(d\omega_{\neg T}) &= \int_{\Omega_{\neg T}} \mathbf{1}_{\left\{\sum_{i \notin T} \mathbf{1}_{\{\omega_i = 1\}} p_i < 2a_n\right\}} \nu_{\neg T}(d\omega_{\neg T}) \\ &= \int_{\Omega_{\neg T}} \mathbf{1}_{\left\{\sum_{i \notin T} \omega_i p_i < 2a_n\right\}} \nu_{\neg T}(d\omega_{\neg T}) \,, \end{aligned}$$

and hence we get

$$\begin{aligned} \int_{\Omega} \mathbf{1}_{\left\{\sum_{i \notin T} \mathbf{1}_{\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\}} p_i < 2a_n\right\}} \nu(d\omega) &= \int_{\Omega} \mathbf{1}_{\left\{\sum_{i \notin T} \omega_i p_i < 2a_n\right\}} \nu(d\omega) \\ &\leq \int_{\Omega} \mathbf{1}_{\left\{\sum_{i=n+1}^{\infty} \omega_i p_i < 2a_n\right\}} \nu(d\omega) \\ &= \int_{\Omega} \mathbf{1}_{\left\{\sum_{i=n+1}^{\infty} \omega_i \hat{p}_i < 2\hat{a}_n\right\}} \nu(d\omega) \\ &\leq e^{-2n} \,, \end{aligned}$$

where the last inequality was established in [13, p. 117]. Hence we find

$$\begin{aligned} & \int_{\Omega} \inf_{n \geq 1} \frac{1}{a_n} \int_{X^n} \mathcal{S}_{\omega,s}(f_{T_n}) \, Q_{\omega}^n(dT) \nu(d\omega) \\ \geq \; & 1 - \sum_{n=1}^{\infty} 3^n \int_{X^n} \int_{\Omega} \mathbf{1}_{\left\{\sum_{i \notin T} \mathbf{1}_{\{\hat{f}_{T_n}(i) \neq 2\omega_i - 1\}} p_i < 2a_n\right\}} \nu(d\omega) \mu^n(dT) \\ \geq \; & 1 - \sum_{n=1}^{\infty} 3^n e^{-2n} \\ = \; & \frac{e^2 - 6}{e^2 - 3} \\ \geq \; & \frac{3}{10} \,. \end{aligned}$$

Therefore, there exists an $\omega \in \Omega$ with

$$\mathbb{E}_{T_n \sim Q^n} \mathcal{S}_{\omega,s}(f_{T_n}) \geq \frac{3a_n}{10}$$

for all $n \geq 1$. □

We now proceed with preparations towards the proof of Theorem 2.2. We begin by recalling the classification framework for DLD introduced in [1]. We have the following definition.

Definition 3.1. Let μ and Q be probability measures on X and $s \in (0,1)$. Then the probability measure $Q \ominus_s \mu$ on $X \times Y$ is defined by

$$Q \ominus_s \mu(A) := s\mathbb{E}_{x\sim Q}\mathbf{1}_A(x,1) + (1-s)\mathbb{E}_{x\sim\mu}\mathbf{1}_A(x,-1)$$

for all measurable subsets $A \subset X \times Y$. Here we used the shorthand $\mathbf{1}_A(x,y) := \mathbf{1}_A((x,y))$ where $\mathbf{1}_A$ is the indicator function of the set A.

Roughly speaking, the distribution $Q \ominus_s \mu$ measures the "1-slice" of $A \subset X \times Y$ by sQ and the "-1-slice" by $(1-s)\mu$. Moreover, the measure $P := Q \ominus_s \mu$ can obviously be associated with a binary classification problem in which positive samples are drawn from sQ and negative samples are drawn from $(1-s)\mu$. Inspired by this interpretation let us recall that the binary classification risk for a measurable function $f : X \to \mathbb{R}$ and a distribution P on $X \times Y$ is defined by

$$\mathcal{R}_P(f) = P\big(\{(x,y) : sign f(x) \neq y\}\big)\,, \tag{3.2}$$

where we define $sign\, t := 1$ if $t > 0$ and $sign\, t = -1$ otherwise. Furthermore, the *Bayes risk* $\mathcal{R}_P^*$ of P is the smallest possible classification risk with respect to P, i.e.

$$\mathcal{R}_P^* := \inf\Big\{\mathcal{R}_P(f) \,\Big|\, f : X \to \mathbb{R} \text{ measurable}\Big\}\,.$$

It is shown in [1] that every distribution $P := Q \ominus_s \mu$ with $dQ := h d\mu$ and $s \in (0,1)$ determines a triple (μ, h, ρ) with $\rho := (1-s)/s$ and vice-versa. We therefore use the shorthand $\mathcal{S}_P(f) := \mathcal{S}_{\mu,h,\rho}(f)$.

In [1] it was shown that $\mathcal{S}_P(f_n) \to 0$ if and only if $\mathcal{R}_P(f_n) \to \mathcal{R}_P^*$. Therefore a classification algorithm which makes $\mathcal{R}_P$ close to $\mathcal{R}_P^*$ also makes $\mathcal{S}_P$ close to zero. Furthermore the following theorem, providing a more quantitative relationship in terms of the ρ-exponent q, was also established.

Theorem 3.1. *Let $\rho > 0$ and μ and Q be probability measures on X such that Q has a density h with respect to μ. For $s := \frac{1}{1+\rho}$ we write $P := Q \ominus_s \mu$. Then the following statements hold:*

(1) If h is bounded then there exists a constant $c > 0$ such that for all measurable $f : X \to \mathbb{R}$ we have

$$\mathcal{R}_P(f) - \mathcal{R}_P^* \leq c\,\mathcal{S}_P(f)\,.$$

(2) If h has ρ-exponent $q \in (0,\infty]$ then there exists a constant $c > 0$ such that for all measurable $f : X \to \mathbb{R}$ we have

$$\mathcal{S}_P(f) \leq c\big(\mathcal{R}_P(f) - \mathcal{R}_P^*\big)^{\frac{q}{1+q}}\,.$$

Theorem 3.1 justifies using learning algorithms designed to minimize the risk function $\mathcal{R}_P$ for the DLD problem. Therefore consider a class of functions $\mathcal{F}$ on X, a loss function $L : \mathcal{F} \times X \times \{-1,1\} \to [0,\infty)$ and denote $L \circ f(x,y) = L(f,x,y)$. Let $P := Q \ominus_s \mu$ where $s := \frac{1}{1+\rho}$. In general, we abuse notation by denoting the empirical distribution corresponding to a sample T with the same symbol T. This

identification implies that the symbol $\mathbb{E}_T$ means the sample average over T. Define an induced loss function

$$L \circledcirc f(x) \ := \ \frac{1}{\rho+1} L(f,x,1) + \frac{\rho}{\rho+1} \mathbb{E}_{x' \sim \mu} L(f,x',-1), \quad x \in X \ .$$

For a training set $T = (x_1, \ldots, x_{n_+}) \in X^{n_+}$ we define the empirical approximation

$$\mathcal{R}_{L,T}(f) \ := \ \mathbb{E}_T L \circledcirc f = \frac{1}{n_+(\rho+1)} \sum_{j=1}^{n_+} L(f,x_j,1) + \frac{\rho}{\rho+1} \mathbb{E}_{x' \sim \mu} L(f,x',-1).$$

Then since

$$\mathcal{R}_{L,P}(f) \ := \ \mathbb{E}_P L \circ f = \mathbb{E}_Q L \circledcirc f,$$

if we can find

$$f_T \in \arg\min \mathcal{R}_{L,T}(f),$$

then the results from [2] could be applied to obtain learning rates. However this approach suffers from the fact that we need to know the functional dependence of $\mathbb{E}_{x' \sim \mu} L(f,x',-1)$ on $f \in \mathcal{F}$. Consequently, in general there appear to be no efficient algorithms for finding f_T. In [2] it was proposed to replace the term $\mathbb{E}_{x' \sim \mu} L(f,x',-1)$ with the empirical approximation $\frac{1}{n_-} \sum_{j=1}^{n_-} L(f,x'_j,-1)$ determined by taking n_- i.i.d. samples $T^- = (x'_1, ..., x'_{n_-})$ from μ. With the appropriate choice of L there exist efficient algorithms but now we are not minimizing a sample average but the convex combination of a sample average over T with a sample average over T^- and so the performance analysis of [2] has to be reconsidered. Instead of analyzing this case in its full generality, here we only consider the case when $n_+ = nm_+$ and $n_- = nm_-$ have a common factor n and derive rates in terms of n. We first explain our approach when $n_+ = n_-$. We define an induced loss function

$$L \circledcirc f(x^+, x^-) \ := \ \frac{1}{\rho+1} L(f,x^+,1) + \frac{\rho}{\rho+1} L(f,x^-,-1), \quad (x^+,x^-) \in X \times X \ .$$

It follows that

$$\mathbb{E}_{Q \otimes \mu} L \circledcirc f = \frac{1}{\rho+1} \mathbb{E}_Q L(f,\cdot,1) + \frac{\rho}{\rho+1} \mathbb{E}_\mu L(f,\cdot,-1) = \mathbb{E}_P L(f,\cdot,\cdot)$$

and

$$\sum_{j=1}^{n} L \circledcirc f(x_j^+, x_j^-) = \frac{1}{\rho+1} \sum_{j=1}^{n} L(f,x_j^+,1) + \frac{\rho}{\rho+1} \sum_{j=1}^{n} L(f,x_j^-,-1)$$

so that the two independent sample averages over X become one sample average over $X \times X$.

Let us now proceed to the more general case $n_+ = nm_+$ and $n_- = nm_-$. Let $L : \mathcal{F} \times X \times \{-1,1\} \to \mathbb{R}$ be a loss function, and denote $L \circ f(x,y) = L(f,x,y)$, $L_+ \circ f(x) = L(f,x,1)$, and $L_- \circ f(x) = L(f,x,-1)$. Let m_+ and m_- be two

positive integers and consider the nm_+-sample $T^+ \in X^{nm_+}$ sampled from Q and the nm_--sample $T^- \in X^{nm_-}$ sampled from μ. Consider the obvious bijections

$$X^{nm_+} \to (X^{m_+})^n$$

$$X^{nm_-} \to (X^{m_-})^n$$

decomposing T^+ into $T^+ = (Z_1^+, ..., Z_n^+)$ and T^- into $T^- = (Z_1^-, ..., Z_n^-)$ with $Z_j^+ \in X^{m_+}, j = 1, .., n$ and $Z_j^- \in X^{m_-}, j = 1, .., n$. They induce a bijection

$$X^{nm_+} \times X^{nm_-} \to \left(X^{m_+} \times X^{m_-}\right)^n.$$

which maps the $nm_+ + nm_-$-sample (T^+, T^-) with points in X to the n-sample

$$\mathsf{T} = \left((Z_1^+, Z_1^-), ..., (Z_n^+, Z_n^-)\right)$$

with points in $X^{m_+} \times X^{m_-}$. Recall that if we consider a point Z^+ in X^{m_+} as an m_+-sample with points in X we write the sample average as $\mathbb{E}_{Z^+}$ and do likewise for Z^- in X^{m_-}. In addition $\mathbb{E}_{\mathsf{T}}$ denotes the sample average over the n-sample T. We now introduce the induced loss function L on $X^{m_+} \times X^{m_-}$ by

$$L \circledcirc f(Z^+, Z^-) \;:=\; \mathbb{E}_{(Z^+ \ominus_s Z^-)} L \circ f = \frac{1}{\rho+1}\mathbb{E}_{Z^+}(L_+ \circ f) + \frac{\rho}{\rho+1}\mathbb{E}_{Z^-}(L_- \circ f). \tag{3.3}$$

The following lemma establishes useful relations for the expected values of sample averages.

Lemma 3.1. *Consider $g : X \times \{-1, 1\} \to \mathbb{R}^+$. For $(T^+, T^-) \in X^{nm_+} \times X^{nm_-}$ decompose T^+ into $T^+ = (Z_1^+, ..., Z_n^+)$ and T^- into $T^- = (Z_1^-, ..., Z_n^-)$ and let $\mathsf{T} = \left((Z_1^+, Z_1^-), ..., (Z_n^+, Z_n^-)\right)$ denote the induced n-sample on $X^{m_+} \times X^{m_-}$. Consider the induced function*

$$\acute{g}(Z^+, Z^-) \;:=\; \mathbb{E}_{(Z^+ \ominus_s Z^-)} g$$

defined on $X^{m_+} \times X^{m_-}$. Then

$$\begin{aligned} \mathbb{E}_{\mathsf{T}} \acute{g} &= \mathbb{E}_{(T^+ \ominus_s T^-)} g, \\ \mathbb{E}_{(Q^{m_+} \otimes \mu^{m_-})} \acute{g} &= \mathbb{E}_{Q \ominus_s \mu} g, \\ \mathbb{E}_{(Q^{m_+} \otimes \mu^{m_-})} \acute{g}^2 &\le \mathbb{E}_{Q \ominus_s \mu} g^2. \end{aligned}$$

Proof. Denote $g_+ \;:=\; g(\cdot, 1)$ and $g_- \;:=\; g(\cdot, -1)$. We have

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_{Z_j^+} g_+ = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{m_+}\sum_{l=1}^{m_+} g_+(Z_{j,l}^+) = \frac{1}{nm_+}\sum_{j=1,l=1}^{n,m_+} g_+(Z_{j,l}^+)$$

and since $Z_{j,l}^+,\ j = 1, n,\ l = 1, m_+$ are the components of the nm_+-sample T^+ the righthand side is equal to $\mathbb{E}_{T^+} g_+$. Therefore

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_{Z_j^+} g_+ = \mathbb{E}_{T^+} g_+ \quad \text{and} \quad \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_{Z_j^-} g_- = \mathbb{E}_{T^-} g_- \tag{3.4}$$

and so

$$\begin{aligned}\mathbb{E}_{\mathsf{T}}\acute{g} &= \frac{1}{n}\sum_{j=1}^{n}\acute{g}(Z_j^+, Z_j^-)\\ &= \frac{1}{n}\sum_{j=1}^{n}\big(\frac{1}{\rho+1}\mathbb{E}_{Z_j^+}g_+ + \frac{\rho}{\rho+1}\mathbb{E}_{Z_j^-}g_-\big)\\ &= \frac{1}{\rho+1}\mathbb{E}_{T^+}g_+ + \frac{\rho}{\rho+1}\mathbb{E}_{T^-}g_-\\ &= \mathbb{E}_{(T^+\ominus_s T^-)}g\end{aligned}$$

establishing the first assertion. For the second assertion observe that

$$\begin{aligned}\mathbb{E}_{(Q^{m_+}\otimes\mu^{m_-})}\acute{g} &= \mathbb{E}_{(Z^+,Z^-)\sim Q^{m_+}\otimes\mu^{m_-}}\Big(\frac{1}{\rho+1}\mathbb{E}_{Z^+}g_+ + \frac{\rho}{\rho+1}\mathbb{E}_{Z^-}g_-\Big)\\ &= \frac{1}{\rho+1}\mathbb{E}_{(Z^+\sim Q^{m_+})}\mathbb{E}_{Z^+}g_+ + \frac{\rho}{\rho+1}\mathbb{E}_{(Z^-\sim\mu^{m_-})}\mathbb{E}_{Z^-}g_-.\end{aligned}$$

Since g is non-negative both g_+ and g_- are non-negative and Tonelli's theorem implies that $\mathbb{E}_{(Z^+\sim Q^{m_+})}\mathbb{E}_{Z^+}g_+ = \mathbb{E}_Q g_+$ and $\mathbb{E}_{(Z^-\sim\mu^{m_-})}\mathbb{E}_{Z^-}g_- = \mathbb{E}_\mu g_-$. Therefore we obtain

$$\mathbb{E}_{(Q^{m_+}\otimes\mu^{m_-})}\acute{g} = \frac{1}{\rho+1}\mathbb{E}_Q g_+ + \frac{\rho}{\rho+1}\mathbb{E}_\mu g_- = \mathbb{E}_{Q\ominus_s\mu}g$$

establishing the second assertion. Finally, Jensen's inequality and Tonelli's theorem imply

$$\begin{aligned}\mathbb{E}_{(Q^{m_+}\otimes\mu^{m_-})}\acute{g}^2 &= \mathbb{E}_{\left((Z^+,Z^-)\sim Q^{m_+}\otimes\mu^{m_-}\right)}\left(\mathbb{E}_{(Z^+\ominus_s Z^-)}g\right)^2\\ &\le \mathbb{E}_{\left((Z^+,Z^-)\sim Q^{m_+}\otimes\mu^{m_-}\right)}\mathbb{E}_{(Z^+\ominus_s Z^-)}g^2\\ &= \mathbb{E}_{Q\ominus_s\mu}g^2.\end{aligned}$$

□

The proof of Theorem 2.2 follows very closely that of the rate theorem for classification using Gaussian kernels [2, Theorem 2.8]. With that in mind we require a slight generalization of [2, Theorem 5.1]. It differs in that it does not require $\mathcal{F}$ to be a set of functions on the domain of the measure space. The proof is essentially the same so it will not be repeated here.

Theorem 3.2. *Let $\acute{P}$ be a probability measure on a set W. Let $\mathcal{F}$ be a convex subset of a vector space and let $L : \mathcal{F}\times W \to [0,\infty)$ be a convex and line-continuous loss function such that the functions $\{L(f,\cdot) : f\in\mathcal{F}\}$ from $W\to[0,\infty)$ are bounded, measurable, and separable with respect to $\|.\|_\infty$. Denote $L\circledcirc f(\cdot) := L(f,\cdot)$ and let $T\in W^n$ be an n-sample. Let $f_{T,\mathcal{F}}\in\arg\min_{f\in\mathcal{F}}\mathbb{E}_T L\circledcirc f$, $f_{\acute{P},\mathcal{F}}\in\arg\min_{f\in\mathcal{F}}\mathbb{E}_{\acute{P}}L\circledcirc f$ and define*

$$\acute{\mathcal{G}} := \left\{L\circledcirc f - L\circledcirc f_{\acute{P},\mathcal{F}} : f\in\mathcal{F}\right\}.$$

and its modulus of continuity

$$\omega_n(\acute{\mathcal{G}},\acute{\varepsilon}) \;:=\; \mathbb{E}_{T\sim\acute{P}^n}\Big(\sup_{\substack{\acute{g}\in\acute{\mathcal{G}},\\ \mathbb{E}_{\acute{P}}\acute{g}^2\le\acute{\varepsilon}}} |\mathbb{E}_{\acute{P}}\acute{g}-\mathbb{E}_T\acute{g}|\Big).$$

Suppose that there are constants $c\ge 0$, $0<\alpha\le 1$, $\delta\ge 0$ *and* $B>0$ *with* $\mathbb{E}_{\acute{P}}\acute{g}^2\le c\,(\mathbb{E}_{\acute{P}}\acute{g})^\alpha+\delta$ *and* $\|\acute{g}\|_\infty\le B$ *for all* $\acute{g}\in\acute{\mathcal{G}}$. *Let* $n\ge 1$, $x>0$ *and* $\varepsilon>0$ *with*

$$\varepsilon\ge 10\max\left\{\omega_n(\acute{\mathcal{G}},c\varepsilon^\alpha+\delta),\sqrt{\frac{\delta x}{n}},\left(\frac{4cx}{n}\right)^{\frac{1}{2-\alpha}},\frac{Bx}{n}\right\}$$

Then we have

$$(\acute{P}^n)^*\Big(T\in W^n:\mathbb{E}_{\acute{P}}L\circledcirc f_{T,\mathcal{F}}<\mathbb{E}_{\acute{P}}L\circledcirc f_{\acute{P},\mathcal{F}}+\varepsilon\Big)\;\ge\;1-e^{-x}.$$

We are now in a position to prove the analogue of [2, Theorem 5.8] when independently the nm_+-sample $T^+\in X^{nm_+}$ is i.i.d. sampled from Q and the nm_--sample $T^-\in X^{nm_-}$ is i.i.d. sampled from μ. Let us introduce some notation. Let $\mathcal{R}_{L,P}(f) := \mathbb{E}_P L\circ f$ denote the L-risk and let $f_{P,\mathcal{F}}\in\arg\min_{f\in\mathcal{F}}\mathcal{R}_{L,P}(f)$ denote a minimizer. For the nm_++nm_--sample $(T^+,T^-)\in X^{nm_+}\times X^{nm_-}$ let $\mathcal{R}_{L,T^+,T^-}(f) := \mathbb{E}_{(T^+\ominus_s T^-)}L\circ f$ and let $f_{T^+,T^-,\mathcal{F}}\in\arg\min_{f\in\mathcal{F}}\mathcal{R}_{L,T^+,T^-}(f)$ denote an empirical minimizer.

Theorem 3.3. *Let* Q *and* μ *be probability measures on a set* X *and let* $P := Q\ominus_s\mu$ *with* $0<s<1$. *Let* $\mathcal{F}$ *be a convex subset of a vector space and let* $L:\mathcal{F}\times X\times\{-1,1\}\to[0,\infty)$ *be a convex and line-continuous loss function such that the functions* $\{L(f,\cdot):f\in\mathcal{F}\}$ *from* $X\times\{-1,1\}\to[0,\infty)$ *are bounded, measurable, and separable with respect to* $\|.\|_\infty$. *Denote* $L\circ f(\cdot) := L(f,\cdot)$ *and consider the class*

$$\mathcal{G} \;:=\; \big\{L\circ f-L\circ f_{P,\mathcal{F}}\;:\;f\in\mathcal{F}\big\}.$$

of functions on $X\times\{-1,1\}$ *Suppose that there are constants* $c\ge 0$, $0<\alpha\le 1$, $\delta\ge 0$ *and* $B>0$ *with* $\mathbb{E}_P g^2\le c\,(\mathbb{E}_P g)^\alpha+\delta$ *and* $\|g\|_\infty\le B$ *for all* $g\in\mathcal{G}$. *Furthermore, assume that there are constants* $a\ge 1$ *and* $0<p<2$ *with*

$$\sup_{(T^+,T^-)\in X^{nm_+}\times X^{nm_-}}\log\mathcal{N}\big(B^{-1}\mathcal{G},\varepsilon,L_2(T^+,T^-)\big)\;\le\;a\varepsilon^{-p} \tag{3.5}$$

for all $\varepsilon>0$. *Then there exists a constant* $c_p>0$ *depending only on* p *such that for all* $n\ge 1$, $m_+\ge 1$, $m_-\ge 1$ *and all* $x>0$ *we have*

$$(Q^{nm_+}\otimes\mu^{nm_-})^*\Big((T^+,T^-):\mathcal{R}_{L,P}(f_{T^+,T^-,\mathcal{F}})>\mathcal{R}_{L,P}(f_{P,\mathcal{F}})+c_p\,\epsilon\Big)\le e^{-x},$$

where

$$\begin{aligned}\epsilon \;:=\; \varepsilon(n,a,B,c,\delta,x) := {}& B^{\frac{2p}{4-2\alpha+\alpha p}}c^{\frac{2-p}{4-2\alpha+\alpha p}}\left(\frac{a}{n}\right)^{\frac{2}{4-2\alpha+\alpha p}}+B^{\frac{p}{2}}\delta^{\frac{2-p}{4}}\left(\frac{a}{n}\right)^{\frac12}\\ &+B\left(\frac{a}{n}\right)^{\frac{2}{2+p}}+\sqrt{\frac{\delta x}{n}}+\left(\frac{cx}{n}\right)^{\frac{1}{2-\alpha}}+\frac{Bx}{n}.\end{aligned}$$

Proof. We intend to apply Theorem 3.2 with $W = X^{m_+} \times X^{m_-}$, measure $\acute{P} = Q^{m_+} \otimes \mu^{m_-}$, and the loss function $L \circledcirc$ defined in (3.3). First observe that Lemma 3.1 implies that $\mathbb{E}_{(Q^{m_+} \otimes \mu^{m_-})} L \circledcirc f = \mathbb{E}_{Q \ominus_s \mu} L \circ f = \mathcal{R}_{L,P}(f)$ and $\mathbb{E}_{\mathsf{T}} L \circledcirc f = \mathbb{E}_{(T^+ \ominus_s T^-)} L \circ f = \mathcal{R}_{L,T^+,T^-}(f)$ so that we obtain the correct risk in the statement of the theorem and the correct empirical risk function to define $f_{T^+,T^-,\mathcal{F}}$. Next we need to translate the variance and supremum bound assumptions on $\mathcal{G}$ to variance and supremum bounds on $\acute{\mathcal{G}}$. To that end observe that for $f \in \mathcal{F}$ the corresponding $g \in \mathcal{G}$ is $L \circ f - L \circ f_{P,\mathcal{F}}$ and the corresponding $\acute{g} \in \acute{\mathcal{G}}$ is

$$\begin{aligned}\acute{g}(Z^+, Z^-) &= L \circledcirc f(Z^+, Z^-) - L \circledcirc f_{P,\mathcal{F}}(Z^+, Z^-) \\ &= \mathbb{E}_{(Z^+ \ominus_s Z^-)}(L \circ f - L \circ f_{P,\mathcal{F}}) \\ &= \mathbb{E}_{(Z^+ \ominus_s Z^-)} g.\end{aligned}$$

Assume for the moment that g and $\acute{g}$ correspond to the same f and so are related in this way. Since $\mathbb{E}_{(Z^+ \ominus_s Z^-)}$ is an averaging operation it follows that $\|\acute{g}\|_\infty \leq B$ if $\|g\|_\infty \leq B$. Moreover, Lemma 3.1 implies that $\mathbb{E}_{(Q^{m_+} \otimes \mu^{m_-})} \acute{g} = \mathbb{E}_P g$ and $\mathbb{E}_{(Q^{m_+} \otimes \mu^{m_-})} \acute{g}^2 \leq \mathbb{E}_P g^2$ so that $\mathbb{E}_{(Q^{m_+} \otimes \mu^{m_-})} \acute{g}^2 \leq c\,(\mathbb{E}_{(Q^{m_+} \otimes \mu^{m_-})} \acute{g})^\alpha + \delta$ if $\mathbb{E}_P g^2 \leq c\,(\mathbb{E}_P g)^\alpha + \delta$. Therefore we can translate variance bounds on $\mathcal{G}$ into variance bounds of the same form on $\acute{\mathcal{G}}$.

We may now apply Theorem 3.2. We need to bound the modulus $\omega_n(\acute{\mathcal{G}}, \varepsilon)$ in terms of the covering bound assumption. Although Jensen's inequality and Lemma 3.1 imply that

$$\omega_n(\acute{\mathcal{G}}, \varepsilon) \geq \omega_n(\mathcal{G}, \varepsilon)$$

this inequality goes the wrong way to be useful. We proceed instead by bounding the modulus of continuity in terms of the local Rademacher average (see [14])

$$\omega_n(\acute{\mathcal{G}}, \varepsilon) \;\leq\; 2Rad(\acute{\mathcal{G}}, n, \varepsilon)$$

and then utilizing [2, Proposition 5.4] to bound the Rademacher average in terms of covering numbers followed by comparing the covering numbers of $\acute{\mathcal{G}}$ in terms of the covering numbers of $\mathcal{G}$. Indeed, write $\acute{g}(Z^+, Z^-) = \mathbb{E}_{(Z^+ \ominus_s Z^-)} g$ and consider an n-sample $\mathsf{T} = \big((Z_1^+, Z_1^-), ..., (Z_n^+, Z_n^-)\big)$ on $X^{m_+} \times X^{m_-}$. Denote $T^+ = (Z_1^+, ..., Z_n^+)$ and $T^- = (Z_1^-, ..., Z_n^-)$. Then by Jensen's inequality we obtain

$$\begin{aligned}\|\acute{g}\|_{L_2(\mathsf{T})}^2 &= \frac{1}{n}\sum_{j=1}^n |\acute{g}(Z_j^+, Z_j^-)|^2 = \frac{1}{n}\sum_{j=1}^n |\mathbb{E}_{(Z^+ \ominus_s Z^-)} g|^2 \leq \frac{1}{n}\sum_{j=1}^n \mathbb{E}_{(Z^+ \ominus_s Z^-)} g^2 \\ &= \mathbb{E}_{(T^+ \ominus_s T^-)} g^2 = \frac{1}{\rho+1}\|g\|_{L_2(T^+)}^2 + \frac{\rho}{\rho+1}\|g\|_{L_2(T^-)}^2\end{aligned}$$

and consequently

$$\|\acute{g}\|_{L_2(\mathsf{T})} \leq \sqrt{\frac{1}{\rho+1}}\|g\|_{L_2(T^+)} + \sqrt{\frac{\rho}{\rho+1}}\|g\|_{L_2(T^-)}.$$

Therefore by choosing a set in $\mathcal{G}$ which determines an ϵ cover for $\|\cdot\|^2_{L_2(T^+)}$ and then for each component of the cover choosing a set in $\mathcal{G}$ which determines an $\epsilon/\sqrt{\rho}$ cover for $\|\cdot\|^2_{L_2(T^-)}$ we obtain that

$$\mathcal{N}\big(\acute{\mathcal{G}}, \epsilon, L_2(\mathsf{T})\big) \leq \mathcal{N}\big(\mathcal{G}, \frac{1}{2}\sqrt{\rho+1}\epsilon, L_2(T^+)\big)\mathcal{N}\big(\mathcal{G}, \frac{1}{2}\sqrt{\frac{\rho+1}{\rho}}\epsilon, L_2(T^-)\big).$$

Consequently assumption (3.5) implies

$$\sup_{\mathsf{T}} \log \mathcal{N}\big(B^{-1}\acute{\mathcal{G}}, \varepsilon, L_2(\mathsf{T})\big) \;\leq\; 8a\varepsilon^{-p}.$$

Thus we can apply [2, Proposition 5.7] to bound the local Rademacher average $Rad(\acute{\mathcal{G}}, n, \varepsilon)$ using $8a$ instead of a. The rest of the proof follows as in the proof of [2, Theorem 5.8] where we use the inequality $\varepsilon(n, 8a, B, c, \delta, x) \leq 8\varepsilon(n, a, B, c, \delta, x)$. □

Proof of Theorem 2.2. The proof follows very closely that of the rate theorem for classification using Gaussian kernels [2, Theorem 2.8]. Let l be the hinge loss defined by $l(y,t) := \max\{0, 1-yt\}$, $y \in Y$, $t \in \mathbb{R}$. We select the loss function

$$L(f, x, y) = \lambda\|f\|^2_H + l(y, f(x))$$

for the no-offset case $\mathcal{F} = H$ and

$$L(f, b, x, y) = \lambda\|f\|^2_H + l(y, f(x) + b)$$

for the offset case $\mathcal{F} = H \times \mathbb{R}$. Consequently for an $nm_+ + nm_-$-sample $(T^+, T^-) \in X^{nm_+} \times X^{nm_-}$, an empirical minimizer $f_{T^+,T^-,\mathcal{F}} \in \arg\min_{f\in\mathcal{F}} \mathcal{R}_{L,T^+,T^-}(f)$ is a solution of equation (2.2) when $\mathcal{F} = H$ and equation (2.3) when $\mathcal{F} = H \times \mathbb{R}$.

A simple calculation shows that the density h has *geometric ρ-exponent α* if and only if $P := Q \ominus_s \mu$ has *geometric exponent α* in the sense of [2]. Moreover [1, Proposition 2.9] shows that h has ρ-exponent q if and only if $P := Q \ominus_s \mu$ has Tsybakov noise exponent q in the sense of [5]. Therefore if we recall that in the proof of Theorem 3.3 we observed that variance bounds on $\mathcal{G}$ imply the same variance bounds on $\acute{\mathcal{G}}$, we conclude that the variance bounds of [2,Proposition 6.1 and Proposition 6.8] hold on $\acute{\mathcal{G}}$ with measure $Q^{m_+} \otimes \mu^{m_-}$. Since we have established Theorem 3.3 as an analogue of [2, Theorem 5.8] and the proof of [2, Lemma 7.2] uses only the function $\epsilon(\cdot)$, [2, Lemma 7.2] holds with $X^{m_+} \times X^{m_-}$ instead of the stated $X \times \{-1, 1\}$. In the same way that [2, Theorem 2.8] follows from [2, Lemma 7.2] we obtain from this modification of [2, Lemma 7.2] that

$$(Q^{nm_+} \otimes \mu^{nm_-})^*\Big((T^+, T^-) : \mathcal{R}_P(f_{T^+,T^-,\lambda_n}) \leq \mathcal{R}^*_P + Cx^2 n^{-\frac{\alpha}{2\alpha+1}+\varepsilon}\Big) \;\geq\; 1 - e^{-x}$$

if $\alpha \leq \frac{q+2}{2q}$ and

$$(Q^{nm_+} \otimes \mu^{nm_-})^*\Big(\mathcal{R}_P(f_{T^+,T^-,\lambda_n}) \leq \mathcal{R}^*_P + Cx^2 n^{-\frac{2\alpha(q+1)}{2\alpha(q+2)+3q+4}+\varepsilon}\Big) \;\geq\; 1 - e^{-x}$$

otherwise. If $\alpha = \infty$ the latter concentration inequality holds if $\sigma_n = \sigma$ is a constant with $\sigma > 2\sqrt{d}$. Furthermore, all results hold for the L1-SVM with offset if $q > 0$. Theorem 2.2 then follows directly from Theorem 3.1 and the inequality $\frac{q}{1+q} \leq 1$. □

References

[1] I. Steinwart, D. Hush and C. Scovel, A classification framework for anomaly detection, *Journal of Machine Learning Research.* **6** (2005), 211–232.

[2] I. Steinwart and C. Scovel, Fast rates for support vector machines using Gaussian kernels, *Annals of Statistics*, submitted, 2004.

[3] I. Steinwart and C. Scovel, Fast rates for support vector machines, to appear *Proceedings of the Conference on Learning Theory (COLT-2005).*

[4] W. Polonik, Measuring mass concentrations and estimating density contour clusters—an excess mass aproach, *Ann. Stat.* **23**(1995), 855–881.

[5] A.B. Tsybakov, On nonparametric estimation of density level sets, *Ann. Stat.* **25**(1997), 948–969.

[6] S. Ben-David and M. Lindenbaum, Learning distributions by their density levels: a paradigm for learning without a teacher, *J. Comput. System Sci.* **55**(1997), 171–182.

[7] I. Steinwart, D. Hush and C. Scovel, Density level detection is classification, *Neural Information Processing Systems* **17**(2005), 1337–1344.

[8] J.A. Hartigan, *Clustering Algorithms*, Wiley, New York, 1975.

[9] A. Cuevas and M. Febrero and R. Fraiman, Cluster analysis: a further approach based on density estimation, *Computat. Statist. Data Anal.* **36**(2001), 441-459.

[10] D.W. Müller and G. Sawitzki, Excess mass estimates and tests for multimodality, *J. Amer. Statist. Assoc.* **86**(1991), 738–746.

[11] G. Sawitzki, The Excess Mass Approach and the Analysis of Multi-Modality, In *From data to knowledge: Theoretical and practical aspects of classification, data analysis and knowledge organization*, 203–211, W. Gaul and D. Pfeifer, Editors, *Proc. 18th Annual Conference of the GfKl*, Springer, 1996.

[12] J.A. Hartigan, Estimation of a convex density contour in 2 dimensions, *J. Amer. Statist. Assoc.* **82**(1987), 267–270.

[13] L. Devroye and L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

[14] A.W. van der Vaart and J.A. Wellner, *Weak convergence and empirical processes*, Springer, New York, 1997.